Use association rules to study the relation between variables that affect high blood pressure

Cirruse Salehnasab[1], Fuad Jahandideh[2], Marzieh Ahmadzadeh[3], Shahram Tahmasebian[*4]

1. School of Management and Medical Information**,** Shiraz University of Medical Sciences, Shiraz, Iran. Email: salehnasab@sums.ac.ir

2. School of Management and Medical Information**,** Shiraz University of Medical Sciences, Shiraz, Iran. Email: fuadjh@gmail.com

3. School of Computer Engineering & IT, Shiraz University of Technology, Shiraz, Iran. Email: ahmadzadeh@sutech.ac.ir

4. Tehran University of Medical Sciences, Tehran, Iran. Email: Tahmasebian@razi.tums.ac.ir

Abstract

Introduction : Due to the increasing development of industrial societies, special diseases were spread in this regard, particularly in Iran, because of improper life style of eating and physical activity, the prevalence of these diseases is high. One of these diseases is high blood pressure, which is the origin of many other diseases and thus, increase costs of the health budget is allocated to it. Usually the types of jobs, lack of exercise and poor diet can have a large impact on the disease.

Methods :In this study, we try to use data mining algorithms; important relation between disease and high blood pressure are effective features, data on 1000 patients who entered our survey.

Results and conclusions: This review was undertaken with association rules employment physical factors and smoking in people with low blood pressure have been seen. Obesity BMI above the low green fruit consumption in people with high blood pressure has been seen together.

Keywords: High blood pressure, data mining, association rules, APRIORI algorithm

*corresponding author

1- Introduction

The pressure of the blood against the walls of the peripheral arteries is caused by the force of the left ventricle. Hypertension is a disease without structural changes in the arteries that can cause the blood to the different parts of the body. Defining "high blood pressure, systolic blood pressure elder than or same to 140 mm Hg, diastolic blood pressure greater than or equal to 90 mm Hg or a means of both".[1] The analysis shows that about 62% of cerebrovascular disease and 49% of heart disease is attributable to poor blood pressure. [2] By understanding the factors affecting, prevention of this disease is possible. Correct the cause of the highest blood pressure, particularly dietary salt intake, and amount of exercise, obesity and excessive drinking of alcohol. And the prevalence of hypertension neutrality is in relation to age, sex, race, and job [2]. Studies suggest a high prevalence of hypertension in Iran. Research at Tehran University in the year 1990showed high diastolic blood pressure in individuals living in Tehran for more than 15 years, systolic blood pressure above about 14% and 18%, respectively. In this study, high diastolic blood pressure in the age group 35 to be 65 years old was found in 47% of men and 55% women [2].From 1990 to 1993, out of the 24 provinces of Iran as health plans and disease, high diastolic blood pressure was 14% of those 14 to 69 years have been reported. In this study, about 34% of people 35 years and older had high diastolic blood pressure in all provinces. [2] The importance of the above factors is affecting hypertension. The aim of this study was to investigate factors associated with non-communicable diseases are the leading cause. Hypertension is among these disorders. Targeting the presence or absence of disease in the population studied by association rules model (APRIORI) to investigate factors affecting blood pressure is discussed. Association Rule is one of the main data mining techniques and almost the most important form of discovery and learning system for extracting patterns (Unsupervised Learning) [6]. They are all interesting patterns and reproducibility and frequent to recover the database. The disadvantage of this method is to extract useful patterns from among all patterns are obtained. In other words, sometimes the

patterns extracted from the database are many useful patterns extracted by an expert will be difficult, and the weaknesses of this method. Two important criteria are the proposed rules include support and confidence. (Base the percentage of transactions that are required by rule, and the transaction is conditional probability confidence that they have rules). In exploring the association rules, rules that have higher support and confidence are important than the others. Task of exploring nature is discovering association rules in large databases using association rules with high confidence and support. APRIORI algorithm of association rules is one of the most popular algorithms. Using this algorithm, patterns and relationships hidden in the data set of association rules are studied.

1-1- APRIORI algorithm

A powerful data mining technique for discovering association elements has led to an Event and APRIORI algorithm is one of the standard methods of association rules. Based iterative algorithm (Frequent) is. During the first set of algorithms for single-element (L1) in the dataset is searched. Repetition is accumulated (if supported by the reader is at least acceptable).L1 and L2 are another element of this search is that it offers. Repeat this procedure to find LK Continues. Database search algorithm is complete. The algorithm has two stages. [7] Firstly, the link collection that contains k elements of Ck Lk-1 with itself, there is a link. Second pruning step: Ck is given by the number of occurrences of each criterion based on the minimum support; the frequencies of omission are less. And Lk is calculated [7] Criterion validity coefficient for selection rules (Confidence) ration Support (Support) further suggests the importance of supporting credit coefficient. (7)

$Support\ (A \rightarrow B) = P\ (A \cup B)$

$Confidence\ (A \rightarrow B) = P\ (A|B)$

2- Previous studies

Using association rules in the field of cardiovascular and also blood pressure - which is a subset of the disease - a study was done, the study resulting from diseases such as high blood pressure which hospital data conducted in 2010 in Korea? This study uses association rules (Model APRIORI, Web Node) on persons 5,022 (2,508 men and 2,514 women) with hypertension was found that two patients associated severe hypertension with diabetes (Support = 35.15 %, Confidence = 100%) and cerebrovascular diseases (Support = 21.21%; Confidence = 100%). (3)

Same study, the APRIORI association rules and models used to determine factors leading to cardiovascular disease, the study groups with a history of stroke, angioplasty, open heart surgery and risk factors of sex (male), smoking, HDL, family history blood glucose, blood pressure history as the most important factor leading to the top three diseases identified. (4)

Other articles in the field of Cardiology as "identify the factors that play a role in heart disease in women and men." Using association rules, reported that the women were less likely to develop heart disease as well as the odds of having sex without signs of angina, as well.

This study used three algorithms (Predictive APRIORI, APRIORI and Tertius). (5) However, the application of data mining techniques has been used areas of cardiovascular diseases, but the use of these methods, focusing on high blood pressure; very little research has been done.

In this paper, using the experience of other's research in cardiovascular diseases to study data mining (Association rules) has been on high blood pressure.

3- Materials and Methods

3-1 – Data selection:

The data in this study, hypertensive disease data collected from non-communicable disease risk factor surveillance Ministry of Health, Treatment and Medical Education of Iran in 2006, is the

provincial Kohgiluyeh and Boyerahmad, the data on 50 clusters in each cluster 10 males and 10 females were considered, in this review also used this way. (Data of 1000 subjects, 500 men and 500 women).Meanwhile, the number of patients with pressure was 233 people. According to the study, high blood pressure, has been considered for the analysis, Pre-processing of data will be conducted as follows:

Variables are used for this study (gender, age, education, ethnicity, occupation, type of oil used, the amount of fruits and vegetables per week, according to minutes of physical activity per week in various conditions, consumption of fish a week, including use of tobacco products (pipe cigarettes, hookah and other tobacco products), BMI (body mass index), waist circumference, and systolic and diastolic blood pressure.

3-2 - Data pre-processing:

Age range, depending on the year of birth and year of data collection in case of subtraction, calculated according to age groups less than 30 years (group 1) and greater than equal to 30 years (Group 2) were divided. This is shown in the first row of Table 1.

Education variable in the nine scale as (1 - the school (illiterate), 2 - less than a fifth elementary, 3 - have completed primary school, 4 - has finished school, 5 - high school is complete, 6 - the course is completed, 7 - is a complete undergraduate, 8 - have completed graduate or above, 9 - does not want to answer or not) in the form 1 - an academic degree 2 - without a college education have the , it causes prevents the scattering data, and show its value in the process of creating a better model. This is shown in the third row of Table 1.

Variable-dimensional, variable ethnicity is in question as a field free to define ethnicity are considered and data indicated that ethnicity Lor frequency 837, Gulf 87 people, leaving 57 people, Arabs = 19 to n = 1 is . Since the cause of this imbalance and balance the effects of the variables in the

model were The first two modes - ethnic Lor, 2 - re-ens were other ethnicities. This is shown in row four of Table 1. The next variable, which is the job description of the job in question was one in the past 12 months (1 - a government employee, 2 - public sector employees, 3 - employed or self-employed, 4 - unpaid work, 5 - Knowledge student or students, 6 - soldier, 7 - house, 8 - retired, 9 - unemployed (able to work), the 10 - unemployed (with disabilities and cannot work), 11 - other professions, 12 - do not know or do not wish to answer) with regard to the distribution and the number of records that this was the way they should be re-end. In his excellent study shows (1 - Employed), 2 - house, 3 - other takes the variable in question. This is shown in the fifth row of Table 1. Another variable, variable type of oil is used for cooking as (1 - Solid vegetable oil, 2 - liquid vegetable oil, 3 - butter or ghee, 4 -, or animal fat, 5 - vegetable butter (margarine), 6 - other, 7 - does not consume any more than the others, 8 - does not use any oil or fat, 9 - Insufficient data (do not know) were d according to the volume of data and efficient, this variable to model them for 1 - 2 oil - more oil was re-ens. This is shown in the sixth row of Table 1.

Another variable in this study, consumption of fruits and vegetables per week basis is Multiplied by the number of units in each of the seven days of the week and finally on fruit and vegetable consumption were gathered together, The variables were d according to the following amount:

1 – It is greater than or equal to 5 2 - less than 5. It is also the seventh row of Table 1 is shown.

The next variable, the variable is the total amount of physical activity in different situations according to the minutes of this variable was calculated And finally for 1 - and 2 more equal 600 minutes per week - less than 600 minutes per week were d. This is the eighth in a row are shown in Table 1.

Another variable is the amount of fish a week for one day - are taking 2 - no use has been end. This is The ninth in a row is shown in Table 1.

Variable rates of consumption of tobacco smoking are as a source of tobacco products for daily use 2 is expressed. It's about the tenth row of Table 1 is shown. Body mass index as a variable that is no longer

divided into two groups: the first group than the second group of 25 were ≥ 25 expression. This is the eleventh in a row of a table is displayed. Another variable is the waist size were divided into two groups, the first group of 102 normal for men and for women less than equal less than equal to 88, the second group are higher than the amounts of fat. This twelfth row is shown in Table 1.

Blood pressure is the last variable, this spread is used to determine a person has high blood pressure, systolic and diastolic blood pressure was measured three times and average it is considered And the index (systolic pressure above 140 mm Hg or diastolic blood pressure above 90 is considered),

Another indicator for having high blood pressure is one of the recommendations of your doctor or health worker 1 - medication that is consumed during the past 2 weeks, 2 - specific diet advice or treatment for weight loss, 3 -, or recommend therapy to quit smoking, 4 - Recommend starting orIncreasing physical activity can be determined is. These variables to either 1 - without hypertension,

2 - High blood pressure have been classified. This is shown in row 1 of Table 4.

Row	Field Name	Value Code	Role
1	AgeCode	1:<30 year 2:>=30 year	Input
2	Sex	1:Male 2:Female	Input
3	Educode	1: COLLEGIATE 2:NON COLLEGIATE	Input
4	EthnCode	1: LOR 2: OTHER	Input
5	JobCode	1: EMPLOYEE 2: HOUSEKEEPER , 3: OTHERS	Input
6	OilUseCode	1: Liquid Oil 2: Other Oil	Input
7	AVFWCode	1: >=5 Unit 2: <5 Unit	Input
8	PHACode	1: >=600min 2: <600min	Input
9	FishConsCode	1: Consuming 2: No Consuming	Input
10	TCCode	1: NO Smoking 2: Smoking	Input
11	BMICode	1:<25 2:>=25	Input
12	WAISTCode	1: Normal(F<=88&M<=102) 2: Fat	Input
13	HBP	1: NO Hypertension 2: Hypertension	Target

Table1: field names and data coding

3-3 - modeling and inference rules:

Analysis of the data was performed using association rules. Thus, both Clementine and Microsoft BI Software are used. It is to be mentioned in these features and is ranked as a whole and not to a specific algorithm.

Value	Field	Rank	
1.0	Sex	1	True
0.99	TC	2	True
0.99	BMI	3	True
0.99	Job	4	True
0.99	AVFW	5	True
0.99	WAIST	6	True
0.99	Age	7	True
0.93	PHA	8	False
0.88	FishCons	9	False
0.19	Ethn	10	False
0.15	Edu	11	False

Table 2: variables priority using feature selection

As seen in Table 2, the highest importance is the Sex field And then TC fields and BMI and Job and AVFW and Age WAIST and also the Sex field are very important field PHA degree fields listed below are important.Finally, FishCons fields and Ethn and Edu not matter in this model. As the field is considered used oil field screen. Fields of study are also consistent with the previously described process is fully described. Next, using Microsoft BI tools, Clementine and that both algorithms are

presented, the rules were extracted as follows: The software Clementine, the field of blood pressure (HBP) as a target or as a result of other variables or conditions leading to the discovery of association rules using APRIORI algorithm we entered. By selecting the minimum confidence (Confidence) = 75% support (Support) = 0.2% were obtained on 28 rolls. Confidences in Table 3 are the rules that will be fully explained later.Microsoft BI in such a manner as to pre-select the HPB field and other fields in the model were as above. Support and Confidence rules were extracted based on the parameters that are listed in Table 4 it will be explained in detail later.

4 - Results, Discussion

4-1-The Extracted Rules-Clementine

As we can be seen in the table below, of the 28 laws of production, 24 percent have four other law's Confidence, Confidence Seventy-five percent of the show.

For example, the law states that the first row is for Workers who have been smoking, physical activity less than 600 minutes a week and has been suffering from abdominal obesity, hypertension is likely to be 100%. Second law states that employees who are tobacco use, physical activities, less than 600 minutes a week and a BMI over 30 with hypertension are likely 100%.

The third row shows that workers who have been smoking, physical activity less than 600 minutes a week have had a hand with abdominal obesity and BMI above 30 are also likely to have been 100% with high blood pressure. The fourth tier law says that workers who are smoking, physical activity less than 600 minutes a week have been hand experience of abdominal obesity, age over 30 years, they also have high blood pressure.

Confidence	Support	Antecedent	Consequent
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and Age = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and OilUse = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and Age = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and OilUse = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and Age = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and OilUse = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and Age = 2 and OilUse = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and Age = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and OilUse = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and Age = 2 and OilUse = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and Age = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and OilUse = 2 and AVFW = 2	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and Age = 2 and OilUse	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and Age = 2 and	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and OilUse = 2 and	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and Age = 2 and OilUse = 2 and	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and BMI = 2 and Age = 2 and OilUse = 2 and AVFW	HBP = 2
100	0.2	Job = 1 and TC = 2 and PHA = 2 and WAIST = 2 and BMI = 2 and Age = 2 and OilUse = 2 and AVFW = 2	HBP = 2
75	0.4	Job = 1 and TC = 2 and WAIST = 2 and FishCons = 2	HBP = 2
75	0.4	Job = 1 and TC = 2 and WAIST = 2 and FishCons = 2 and BMI = 2	HBP = 2
75	0.4	Job = 1 and TC = 2 and WAIST = 2 and FishCons = 2 and Age = 2	HBP = 2
75	0.4	Job = 1 and TC = 2 and WAIST = 2 and FishCons = 2 and BMI = 2 and Age = 2	HBP = 2

Table 3: Rules extracted with confidence> %75

Similarly, other rules can be seen from Table 3. Extracted according to the roll of tobacco workers in all 28 have been recorded. The next level of physical activity at least sees the people in the 24 record are observed. On the other hand, those three factors have abdominal obesity were also hypertensive patients. Use oil other than oil, low green fruit, fish consumption, and age not over 30 years of other factors lead to high blood pressure. Although each of these factors independently, the following clinically proven that Influence on high blood pressure, but keeping these factors in this study can be unremarkable.

4-2-The Extracted Rules-Microsoft BI

From total of 360 achieved in software rules, the rules with probability greater than or equal to 40% have been considered. For example, the first row of people who have good green fruits and non-Lor race is likely to have hypertension 66.7.

The person in the second row right green fruits with low physical activity on the risk of hypertension had 66.7. Third row right person is working with green fruits with a 50 percent chance of becoming infected. Fourth row right green fruit consumption with age over 30 years and have a 48% chance. Similarly, other laws can be seen from Table 4.

Confidence %	Support %	Antecedent	Consequent
66 7	41	AVFW = 1 and Ethn = 2	HBP = 2
66 7	39	AVFW = 1and PHA = 2	HBP = 2
50	31	AVFW = 1 and Job = 1	HBP = 2
48	31	AVFW = 1 and Age = 2	HBP = 2
44	28	TC = 2 and WAIST = 2	HBP = 2
43	35	WAIST = 2 and Fish Cons = 2	HBP = 2
43	26	AVFW = 1 and BMI = 2	HBP = 2
42	26	AVFW = 1 and Edu = 1	HBP = 2
42	43	WAIST = 2 and Age = 2	HBP = 2
40	23	AVFW = 1 and WAIST = 2	HBP = 2
40	23	WAIST = 2 and Sex = 1	HBP = 2
40	23	Oil Use = 1 and Job = 2	HBP = 2

Table 4: The rule extracted for hypertension

Rules can be seen in Table 4 that rather the age, the major risk factors for high blood pressure, poor diet, especially consumption of fruits and vegetables is low. In this table, the proper use of green fruit is seen in patients with high blood pressure, which can be attributed to the improved food pattern after infection. The rules are overweight, smoking, low education over 30 years of age in hypertensive subjects shows.

5 -Conclusion

In this study, we use patient data to assess causes affecting non-communicable disease risk factors on disease associated with high blood pressure (which is one of the most common non-communicable diseases).

This review was undertaken with association rules employment physical factors and smoking in people with low blood pressure have been seen. Obesity BMI above the low green fruit consumption and animal's fats in people with high blood pressure has been seen together. Another common age is30.

6-Acknowledgment

At the end, we appreciate the respectful employees of the management of the diseases which are non-communicable, the sanitary deputy ship of Yasuj University of Medical Sciences because of sincere cooperation and guiding what led to codification and presenting the essay.

7-References

1. Mark H. Beers."The Merck Manual of Diagnosis and Therapy". Edition, 2006.John Wiley & Sons

2. Health, M. o. H. a. M. E.-D. o. (2005). "Non-communicable disease risk factor surveillance system."

3. Shin Am Fau - Lee, I. H., G. H. Lee Ih Fau - Lee, et al. "Diagnostic analysis of patients with essential hypertension using association rule mining." (2093-369X (Electronic)).

4. Karaolis, M., J. A. Moutiris, et al. (2009). Association rule analysis for the assessment of the risk of coronary heart events. Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.

5. Nahar, J., T. Imam, et al. (2013). "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications **40**(4): 1086-1093.

6. T.M., M. (1999). "Machine Learning and Data Mining." Communication of the ACM **42**(11).

7. Jiawei Han, M. K. (2006). "Mining: Concepts and Techniques, Second Edition "Elsevier.